·方法学研究·

基于机器学习技术挖掘中医名家医案数据的方法探讨

夏 鑫[1]，牟 玮[2]，李艳芬[2]，黄宇虹[2]

1. 天津中医药大学研究生院（天津 301617）
2. 天津中医药大学第二附属医院临床药理中心（天津 300150）

【摘要】名家医案作为解决临床疑难问题的典范，其诊疗思路的正确性和实践结果的有效性在长期的临床活动中得到了肯定。但传统的统计方法在面对中医学非线性、多维度、复杂关系的问题上，难以全面且深入地发掘中医学的思辨体系和经验逻辑，而机器学习方法在此类问题研究上有显著优势，并在中医传承研究中得到了广泛的应用。本文旨在探讨如何利用机器学习算法研究中医医案，并介绍了中医医案数据的获取与处理、机器学习算法的选择等，以期为中医医案的研究提供参考。

【关键词】中医；医案；数据挖掘；机器学习

Approaches to the mining of traditional Chinese medical experts' case histories using machine learning techniques

XIA Xin[1], MU Wei[2], LI Yanfen[2], HUANG Yuhong[2]

1. Graduate School, Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China
2. Department of Clinical Pharmacology, the Second Affiliated Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin 300150, China
Corresponding author: LI Yanfen, Email: Liyanfen2008@163.com; HUANG Yuhong, Email: hyh101@126.com

【Abstract】As a model for solving clinical challenges, renowned medical cases have affirmed the correctness of diagnostic and treatment approaches as well as the effectiveness of practical outcomes through long-term clinical practice. However, traditional statistical methods struggle to comprehensively and deeply unveil the speculative system and empirical logic of traditional Chinese medicine (TCM), especially when confronted with its nonlinear, multidimensional, and complex relationships. In contrast, machine learning methods demonstrate significant advantages in addressing such issues and have been widely applied in the study and inheritance of TCM. This article aims to discuss how to use machine learning algorithms to study TCM medical cases, and to describe the acquisition and processing of TCM medical case data and the selection of machine learning algorithms, with a view to providing references for the study of TCM medical cases.

DOI: 10.12173/j.issn.1004-5511.202312129
基金项目：国家重点研发计划“中医药现代化研究”专项（2021YFC1712900、2021YFC1712902）
通信作者：李艳芬，博士，主任医师，硕士研究生导师，Email：Liyanfen2008@163.com
黄宇虹，博士，研究员，博士研究生导师，Email：hyh101@126.com

【Keywords】Traditional Chinese medicine; Case histories; Data mining; Machine learning

中医药学是中华民族的伟大创造，是中国古代科学的瑰宝[1]。总结和传承中医名家的辨证思路、治法治则、用药规律等临床经验，通过数据挖掘技术提炼学术思想并予以推广，对丰富中医学理论体系，推动中医药发展意义重大而深远。

传统的医案群案研究主要以统计归纳方法为主，该方法适用于中医医案中病症、治疗方法等信息的频数统计、相关分析，对于揭示特点和规律有一定适用性，但在处理复杂关系和个体差异性方面存在局限，如难以处理多个因素的相互影响，以及多维非线性关系等复杂问题。随着人工智能技术的不断成熟，以及多学科交叉融合的研究和应用，机器学习方法作为数据挖掘技术的基本形式，有望在中医传承中得到广泛应用。中医理论和临床应用来源于长期大量的实践活动，而机器学习算法也是通过分析和学习数据总结规律和预测结果。从临床中发掘中医诊治规律更适合中医理论特点[2]，机器学习算法可从大量数据中挖掘出中医证候特征和治疗用药规律，全面真实地发掘与传承中医名家的思辨体系和经验逻辑。

机器学习基于数学的理论和方法来构建和训练模型，实现对数据的学习和预测。利用机器学习方法将中医技术进一步标准化、形式化，既有助于中医技术的稳定传承和解析研究，又能推动中医的智能化发展。但是由于机器学习技术算法众多，不同机器学习算法都有其特点、适应和限制。挖掘名中医经验传承数据时，数据处理与获取以及算法选择对挖掘医案规律和构建名家诊治模型至关重要。本文对基于机器学习技术挖掘中医名家医案数据的不同方法进行探讨，为中医药相关研究人员提供一定参考。

1 医案数据的获取和预处理

1.1 名家医案数据价值

机器学习的基础是数据，高质量临床数据的获得对于数据规律的挖掘和机器模型的训练至关重要。尤其是面对临床真实情况的复杂多样，出于假设空间的满足，对数据量提出了更高的要求[3]。高质量且数量充足的数据，有助于提高模型的准确性和特殊情况的可靠性。传统中医主要通过四诊“望、闻、问、切”的方式采集信息，缺乏临床的客观指标和一致性标准，对于辨证诊断存在较大的主观性。加之中医关于中医的疗效判定问题，一直是个存在争议且未能得到根本解决的问题[4]。而名家医案作为解决临床疑难问题的典范，其诊断治疗思路和实践结果的有效性在长期的临床活动中得到了肯定，适宜作为用于机器学习的高质量数据来源。

1.2 文本数据结构化

中医文献和病案大部分以文本保存，数据积累较多，语言表述复杂。由于中医医案多为非结构化文本，机器难以直接处理，需要转化成高度组织和整齐格式化的数据，常见方式是将其转换为表格形式。机器学习需从案例中提取有意义的信息特征，如患者的性别、年龄、体质、症状体征、舌脉象等临床信息，以及中医诊断、治疗方法、具体用药等信息。既往研究已证实，自然语言处理技术和大语言模型可用于非结构化文本的提取。文天才和李平开发了采用标引技术将医案结构化的系统，为数据提取提供了有力支持[5]。邓宇等人利用信息抽取方法，对中医医案术语抽取方法进行初步探索，其中遗漏或错误术语约占6%，错误原因主要包括语料库未涵盖所需术语词汇、疾病名称与症状术语之间的混淆、子句分割错误，以及原文中标点符号使用不准确等问题[6]。中医医案主观性强，结构繁杂，内涵外延模糊，形式多样[7]，不同医家对同样症状的理解和描述可能不完全相同。加之目前尚未形成全面统一的中医术语标准，对于多个中医术语的名称和定义仍未达成一致理解[8]。此外，中医医案往往还包含有隐形信息，隐性知识的显性化对于名老中医经验的传承工作有重要意义[9]。虽然有自然语言和文本挖掘技术可以辅助处理，但是出于上述原因，以及医疗活动要求的精确严谨，目前中医医案信息获取还非常依赖专家经验。人工提取数据可更好地理解医案的语言，更准确地提取有价值的信息，同时也是中医学习的重要途径。经过中医学专业人员规范提取的数据，可以更简便地应用于其他学科。

1.3 数据预处理

数据预处理是机器学习关键步骤之一，主要工作有数据清洗和数据规范。数据清洗包括去除数据中重复数据、噪声，修复错误数据，处理缺失值等。数据规范化是中医医案研究非常关键的一部分，由于中医语言的丰富性和长期的历史演变，中医文本数据具备抽象和不规范的语言特点，通常存在着一些词语的同义、近义、一词多义以及模糊、省略和指代等表达方式[10]。这样的数据如果不归纳到同一个标准描述下，将导致数据集维度过大且变得十分稀疏。中医药行业术语的标准化对于消除语义模糊、统一定义至关重要[11]，可显著提高中医药研究的准确性和可靠性，编制中医同义词字典在这个过程中扮演着重要的作用[12]。此外，中药的规范问题有待解决。除了异名外，一是中药存在生熟异用，中药经炮制后能够减毒增效，饮片“生熟异用”是中医临床有效性与安全性的重要保证[13]，对于不同炮制药物是否进行合并统一，需要临床大夫的经验和判断；二是对于剂量缺乏统一定量，所谓“中医不传之秘在药量”，中国古代度量衡制度在不同朝代存在差异，且存在“枚”“个”“粒”“鸡子大”“弹丸大”“把”“握”等非计量标准，利用知识和经验判断对于它们的规范化转化也是不能忽略的重要问题。通过数据清洗和数据规范化，可以有效解决中医文本数据中的各种问题，为后续的机器学习分析和名医模型提供高质量的数据。

1.4 特征选择

在数据预处理的基础上，进行特征选择是进一步优化数据集的重要步骤。特征选择旨在从众多特征中挑选出对数据分析和名医模型训练有重要贡献的特征，可以降低数据维度、提高模型性能等。对于特征的选择，可以根据统计学的相关系数、卡方分析等结果，结合中医领域知识，以及数据分析结果和模型表现性能综合评估。

2 机器学习方法的选择

机器学习专注于使用数据和算法，旨在构建能够根据所使用的数据进行学习的系统。机器学习方法可以高效、准确、统一处理大量的数据，尽可能避免人工分析中的主观性和误差。该方法既可以对诊断、治疗等数据进行全面的分析，挖掘其中的隐藏数据规律，更能构建中医认知模型，实现机器对疾病的诊断和治疗，推动中医学的传承应用和创新发展。根据数据是否存在人工标记，机器学习算法主要分为无监督学习算法和监督学习算法。

2.1 无监督学习

无监督学习是在没有标注数据的情况下，通过算法自动发现数据中的模式和规律。因为它不依赖预设的标签，在一定程度上排除了主观因素的影响。传统中医诊断和治疗基于患者的主观感受和医生的主观判断，结果的准确性与医生个人的能力和经验直接相关[14]。这种主观因素使中医自身发展的途径有着充分的开放性和发散性，是中医自身发展的源泉，但同时也造成了中医传承的障碍，导致了中医医生整个群体的良莠不齐，一定程度上限制了中医发展[15]。无监督学习通过数据本身的特征和规律结构，通常用于探索性目的。在医案数据挖掘中，无监督学习主要用于聚类和降维。

2.1.1 聚类算法

聚类分析是根据数据的相似性和相关性差异将其分为不同的群组。辨证论治是中医诊断和治疗疾病的基本原则，中医基于“整体观念”和“天人合一”的思想与“四诊合参”的诊断依据，其疾病分型依据复杂，不适宜单一暴露因素的队列研究[16]。聚类分析以中医四诊信息为特征，将相似患者分为一组，结合中医知识的分析，从而帮助中医医生更好地辨证分型。也可以中药处方为特征，将相似中药处方分为一组，结合中医知识的分析，从而帮助中医医生挖掘药物组合规律、发现新的药物组合等。

常用的聚类算法主要是 K 均值聚类和层次聚类。K 均值聚类根据设定的聚类数量分组数据，需要对聚类数量有合理的估计或选择，可以结合既往知识以及肘部法则、轮廓系数等评估方法，验证和调整聚类数量。存在对初始质心敏感的问题，适用于数据具有明显的簇结构和较为均匀分布的情况。张佳康利用 K 均值聚类算法，利用知网数据库的 239 个相关中药处方，以每个案例为样本，所使用的具体中药为特征，得出多个中药聚类，找到了气郁痰阻型甲状腺结节的核心药物组成，并依据此核心方药组成结合专家经验，自拟消瘿汤，在临床上有良好疗效[17]。层次聚类通过逐步分割或者合并来形成聚类结构的层次，层

次聚类的结构视图可灵活直观判断聚类数量，非常适用于探索性数据分析和可视化。庄逸洋等利用古今医案云平台的层次聚类算法，对邓铁涛治疗冠心病的用药进行了挖掘和研究，找出了用药核心组合和候选新处方[18]。

2.1.2 主成分分析

主成分分析是可以降低数据维度的方法，其可去除冗余信息，提取数据的有用结构信息，从而更好地理解和分析数据。从医案诸多症状体征中抽取主要特征进行降维，排除不相干的症状，使中医辨证论治，化繁为简，在临床上具有重要意义，有助于中医医生更好地理解疾病的特点，为辨证施治提供指导。刘静等对 300 例不同中医证型腰椎间盘突出患者的四诊信息经主成分分析，以特征值、方差贡献和旋转后因子负荷矩阵结果，得出不同证型的主症和次症，并根据中医理论分为不同证型[19]。李毅同样利用主成分分析结果，结合贡献率、特征根和专业知识，获得溃疡性结肠炎症候的主要症状指标，该结果与中医临床辨证基本相符[20]。谭楠楠等运用主成分分析，以慢性心力衰竭的复杂临床症状为特征，将纳入的症状进行降维，得到了多个症候群，降维后的症状能够明显地进行证候的判断，并对结果分析发现“痰”有最高的症状累计贡献率，肯定了在心力衰竭中化痰逐饮法的认识[21]。

2.1.3 因子分析

因子分析也是可以降低数据维度的方法，其从多个变量中提取出更少的具有特征代表性的几个潜在公因子，能更好地消除冗余信息，简化分析。徐小港等以痛经使用的高频中药进行 KMO 检验和球形检验后，表明该数据具备因子分析意义，经因子分析后找到 5 个累积方差贡献率较高的公因子，公因子内的用药组合结合关联规则结果以及中医分析，确保了结果的可靠性[22]。史媛媛运用因子分析法，分别对不同历史时期的不寐患者医案的高频症状、舌脉象以及治疗用高频中药进行分析，结果可以简易体现出明清和现代在不寐诊治上的差异[23]。

2.2 监督学习

监督学习利用已有的标记数据训练模型，并对未知的数据进行预测。在医案数据挖掘中，监督学习主要用于分类问题。名家医案通常被认为在诊断和用药方面具有较高的准确性和一定的普适性，但同时也应充分考虑不同学派的医家可能存在的认知差异，可依据不同学派数据源建立相应学派模型，并且以集成学习方式将多个模型融合（表 1）。在特征工程阶段，引入医家标识的特征，以确保模型能够反映不同医家之间的差异性。学习名家医案的诊治规律，可以实现机器对疾病的诊治。相比于无监督学习，监督学习有着明确的训练指标，即模型在测试集上的性能。案群研究主要以分类问题为主，该类问题的衡量指标往往为准确率、曲线下面积（area under the curve，AUC）值、F1 得分等[24]。准确率反映了正确预测的样本比例，但可能受数据集不平衡影响；AUC 衡量模型对正负样本的整体区分能力，值越接近 1 表示性能越好，越接近 0.5，则表示模型效果接近于随机猜测；F1 得分是精确率和召回率的调和平均数，综合评估模型的精确性和完整性。这些指标帮助研究者全面评估模型的分类效果。

常用的监督学习算法包括以下几种：决策树和随机森林算法、支持向量机算法、贝叶斯算法、Logistic 回归分析等。

2.2.1 决策树和随机森林算法

决策树算法解决分类问题并用树状图的结构呈现，能清晰地解释和理解决策过程，并对缺失值和异常值具有较好的容忍性。但在处理大样本集时，决策树易出现过拟合现象，从而降低分类的准确性，可对决策树进行剪枝，从而简化决策过程。决策树在中医领域有广泛应用，对复杂的诊断数据，余学杰等采用 ID3 决策树算法，以中医四诊为特征、专家辨证为标签，通过信息增益率和决策树树形结构，理清了证候与证名之间的关系，确认了决策树方法的可行性[25]。刘广采用决策树 C4.5 算法对 800 名胃炎中医患者进行辨证法则的学习，提炼出决策树所包含的分类法则[26]。黄嘉韵对 560 例鼻部患者采用 CART 决策树算法构建证型决策树模型，利用症状判断证型，辨证准确率达 91.5%[27]。苏翀等以中医四诊信息为特征、以是否为慢性阻塞性肺病为标签，分别使用 ID3、C4.5、CART 和 KLDDT 建立多种决策树诊断模型，并且考虑误诊的代价，采用多种衡量指标，得出 KLDDT 决策树有更少的漏诊和误诊率，更适合作为疾病诊断模型[28]。除了剪枝避免过拟合外，采用由多个决策树的投票机制构建的随机森林，也能有效防止过拟合，通常也比单一的决

策树具有更好的性能。宫文浩等利用随机森林的方法，建立了小儿肺炎痰热肺病的诊断辨证模型，取得了较好的分类效果[29]。姜超等利用随机森林、人工神经网络、贝叶斯网络的方法，对 518 例脑出血急性期患者的中医四诊信息建立脑出血中经络和中脏腑辨证模型，结果显示随机森林分类效果最好[30]。由于随机森林集成了多个决策树，其可解释性较决策树相对差，通常需要查看每棵树的平均贡献或特征重要性来理解随机森林。

2.2.2 贝叶斯算法

朴素贝叶斯算法以概率推理为基础，擅长对复杂不确定性、关联性导致的问题进行求解。贝叶斯网络充分考虑了特征之间的相互关系，并使用因果推理预测可能性。在中医证素辨证体系中，运用贝叶斯网络研究症状与证素间的隶属关系以及证素之间的组合关系，所得结果与中医专家的经验表现出高度的一致性[31]。双相情感障碍抑郁发作特征主观且躯体、精神症状繁多，并且包含其他中医四诊信息，特征复杂，刘鑫子等使用基于约束学习方法中的 PC（Peter & Clark）算法，利用贝叶斯网络模型寻找关系，将较强联系的证素结合再经人工判定，得到了证素之间的联系并使研究更加客观[32]。甘小金运用贝叶斯网络方法做分类识别，将王子瑜教授的 150 例医案，以症状特征、证素信息为标签，通过相关性分析筛选出对于证候要素有意义的症候群作为先验模型，使用交叉验证评估性能并获得特征贡献度，该模型对证素判断有良好性能，认为该算法在传承老专家学术经验方面具有较好的应用前景[33]。

2.2.3 支持向量机算法

支持向量机是一种以寻找超平面来分割样本为目的的二类分类模型。在中医的实际应用中，一般要解决多类的分类问题，这时可以考虑通过构建多个二类分类器组合来解决[34]。中医症状信息具有非线性和多维性的特点，支持向量机使用非线性映射，在高维空间获得数据之间更复杂的非线性关系。许明东等人采用支持向量机算法，以高血压病常见的症状及舌脉象为特征，构建了 5 个二分类的支持向量机模型并对 5 个中医证型分别预测，结果总体准确率为 90.0%，认为该算法具有较高的证型诊断准确性[35]。顾天宇等人以中风病患者的一般信息和该病常见症状和舌脉象为特征，以是否为气虚血瘀证为标签，建立支持向量机、BP 神经网络、梯度提升决策三个模型，综合对比下认为对核函数进行调参的支持向量机模型性能优于另外两个模型[36]。

2.2.4 Logistic回归分析

Logistic 回归分析将一个或多个自变量与一个二分类的因变量相关联，尽管该算法被称为“回归”，但它实际上是一种用于处理分类问题的算法，具有良好的解释性。王伟杰等以中医症候为特征，分别以某个中医证型判断的是否建立多个二分类非条件 Logistic 回归模型，并逐步引入变量，分析出了对每个证候辨证有决定意义的主要症状[37]。仲芳在对处方药物剂量统一规范化处理后再对该数据进行分箱处理，以药物及其剂量区间为自变量，治疗结果为因变量，对样本较多的脉弦细组采用 Logistic 回归分析方法，挖掘药物在不同剂量时对于症状愈后的影响，初步明确了药物有效区间和较为明确的量效关系，为临床精准用药提供了参考，对于中药剂量数据规范化、标准化的预处理，该文章也提供了一定参考[38]。

2.3 常用机器学习方法的特点

2.3.1 无监督学习的局限性

无监督学习旨在揭示数据的内在结构和模式，可以为关联分析、特征提取等提供有价值的信息，对于辅助研究和临床实践有一定的参考意义。然而，无监督学习算法只能利用患者病情和治疗中的内在结构分析，并不能对数据特征本身的权重有特别认识。病有轻重缓急，药有君臣佐使，分析结果并不一定完全符合临床的实际应用，因此无监督学习必须依赖中医医师对结果的解读，才能做出对临床有意义的解释和说明，对于无监督学习的结果需要审慎看待。

2.3.2 不同监督学习算法的特点

决策树在可解释性方面具有优势，随机森林牺牲解释性获得了更好的性能，支持向量机算法在非线性、高维度的小样本分类问题上有突出优势，贝叶斯网络基于概率对不精准、不完全、模糊的不确定性问题有良好表现[39]，而 Logistic 回归分析算法在解释性和处理二分类问题方面较为突出。相比于传统的推论统计，机器学习算法通常不拘泥于特定的假设，通过数据之间存在的模式和规律进行预测，并且可以使用多种模型和算法进行建模和预测。因此监督学习算法往往不拘泥于特定的选择，而是根据最终表现结果进行选

择。结合多种算法优势互补、相互交融、综合分析，对模型的性能表现和结果的解释具有重要意义。

在医学研究中，推荐尽可能采用可靠和可解释性模型，以便理解预测模型，对结果进行深入分析。支持向量机算法难以直观地理解模型的决策过程，其可解释性相对较低。贝叶斯算法模型的结果更偏向于概率分布的表达，而不是直接的决策过程。Logistic 回归分析算法通过回归系数来解释特征对结果的影响程度，具有较好的可解释性，然而，它在处理高维数据和非线性关系时的表现较弱。决策树算法作为一种透明的模型，在可解释性以及多种问题的处理方面具有优势，与人脑思维有较高契合度，可以作为医案学习的优先考虑（表 1）。

表1 常用机器学习算法特点

Table 1. Features of commonly used machine learning algorithms

常用机器学习算法	优点与适应性	缺点与局限性
无监督学习（无标注，探索医案规律）		
聚类算法	根据病例的相似性分组；可用于症候、中药等分组	需要事先确定分组的个数；评估聚类质量和分析结果困难
主成分分析	降低数据复杂性，简化数据分析；提取症候组、中药组重要特征	对中医研究的非线性关系难以捕捉；次要信息丢失
因子分析	降低数据复杂性，简化数据分析；解释性较强，方便理解；识别潜在症候组或用药组	对中医研究的非线性关系难以捕捉
监督学习（有标注，建立名医模型）		
决策树算法和随机森林	对特征的缺失不敏感；非线性关系的捕捉，在中医数据上有更好适用性；决策树模型的解释性强，容易理解	对特征选择敏感
支持向量机算法	在小样本上表现良好；有效处理非线性问题和多特征的高维数据	对参数设置敏感，计算复杂度高
贝叶斯网络算法	概率推理，揭示复杂关系；直观的图形化，易于理解；结合中医先验知识	依赖中医先验知识的可靠性
Logistic回归分析	适用于二分类问题；模型的解释性强，易于理解，实现简单	对特征的选择敏感；对特征相关的共线性问题较为敏感

3 结语

机器学习算法对于中医医案数据特点有良好的适应性，无监督学习在提供辨证依据、总结辨证特点、寻找核心药物组成和新处方等挖掘数据内部特征上有广泛应用。监督学习在学习名老中医经验，实现决策的自动化上有良好表现，其中，决策树算法更能可视化体现决策过程，获取专家辨证规律，为医师提供新的思考线索。

目前对古代名家医案的研究，大部分停留在数据规律的挖掘上，对于监督学习模型的建立较少。这与前文所述的中医医案的复杂性和模糊性有很大的相关性。因此本文具体算法的举例多以非名医医案为研究对象。这些医案也主要以中医四诊为特征，与医案蕴含的信息相同，具有一定的参考价值。

获取优质的中医医案数据，合理选择和提取特征，选择合适的机器学习算法展开研究，可更好地推动中医诊断和治疗的研究。目前机器学习在自然语言处理上得益于大型语言模型的发展，已经取得了巨大的进步。以 ChatGPT 为代表的语言模型实现了显著的性能提升，并且具备了指令遵循的能力[40]，可以通过自然语言微调模型，为医案的学习提供新的可能。但是大模型的幻觉问题，即生成虚构或误导性的信息，以及其黑箱特性导致的解释困难，是其应用于医疗领域中不可忽略的问题[41]。传统的机器学习模型高度依赖数据规模，对于医案数量和质量有一定要求。迁移学习可以通过将某一相关领域学习到的模式应用于当前领域来解决上述问题，并构建中医认知的源模型[42]，可为研究其他中医目标领域提供初步的模型参数、特征表示或知识，从而加速目标

领域的学习过程并提高模型的泛化能力，实现机器学习的“举一反三”。因此，结合中医理论和经验获得的先验知识，可进一步指导模型的建立和优化，使其更加贴合中医认知的特点和规律。现行的机器学习标注和研究结果对于中医认知源模型的构建具有重要意义。通过分析和学习这些结果，可以获得更好的模型参数和特征表示，进而提高模型在中医认知方面的预测和推理能力。除了在自然语言上的成功，机器学习在图像识别和分类上也有着显著的进步，较之古代名医，当代名医可以保留原始图像作为机器学习的研究对象，对中医望诊的准确把握和研究提供了更客观的方法。随着机器学习在中医方面的不断研究，中医药的传承和发展将在这条路上迎来新的机遇和挑战。

参考文献

1 肖圣鹏，崔友平 . 坚定中医药自信发展中医药事业 [J]. 红旗文稿, 2019, (16): 34–35. [Xiao SP, Cui YP. Strengthen confidence in traditional Chinese medicine and develop the cause of traditional Chinese medicine [J]. Red Flag Manuscript, 2019, (16): 34–35.] DOI: CNKI:SUN:HQWG.0. 2019–16–012.

2 代倩倩，王燕平，商洪才，等 . 从循证医学与转化医学谈中医药临床研究发展 [J]. 生物医学转化，2022, 3(3): 2–6. [Dai QQ, Wang YP, Shang HC, et al. Discussion on the development of traditional Chinese medicine clinical research from evidence–based medicine and translational medicine[J]. Biomedical Transformation, 2022, 3(3): 2–6.] DOI: 10.12287/j.issn.2096–8965.20220301.

3 Jennings NR，Wooldridge MJ. Foundations of Machine Learning[M]. MIT Press, 2012.

4 庄铭，安佳丽，钟梦媛，等 . 中医药临床疗效评价方法研究进展 [J]. 中国中药杂志，2023, 48(12): 3263–3268. [Zhuang M, An JL, Zhong MY, et al. Evaluation methods of clinical efficacy of traditional Chinese medicine[J]. China Journal of Chinese Materia Medica, 2023, 48(12): 3263–3268.] DOI: 10.19540/j.cnki.cjcmm.20230219.502.

5 文天才，李平 . 基于 XML 的名老中医医案结构化标引系统 [J]. 中国数字医学，2013, 8(7): 22–24. [Wen TC, Li P, The structuring index system for famous TCM doctors medical record based on XML[J]. China Digital Medicine, 2013, 8(7): 22–24.] DOI: 10.3969/j.issn.1673–7571.2013.07.006.

6 邓宇，张振铭，陈橙，等 . 基于正则表达式的中医医案术语抽取方法研究 [J]. 湖南中医杂志，2023, 39(5): 202–207. [Deng Y, Zhang ZM, Chen C, et al. Research on the terminology extraction method of traditional Chinese medicine medical case based on regular expressions[J]. Hunan Journal of Traditional Chinese Medicine, 2023, 39(5): 202–207.] DOI: 10.16808/j.cnki.issn1003–7705.2023.05.045.

7 谢蓉，王燕萍，彭丹虹，等 . 中医症状规范化研究 [J]. 河南中医, 2017, 37(7): 1144–1146. [Xie R, Wang YP, Peng DH, et al. The research of the standardization of TCM symptoms[J]. Henan Traditional Chinese Medicine, 2017, 37(7): 1144–1146.] DOI: 10.16367/j.issn.1003–5028.2017.07.0403.

8 吴文玲 . 面向中医诊疗知识库的术语规范化研究 [D]. 长春：吉林大学，2021. [Wu WL. Study on the standardization of terms in tcm diagnosis and treatment knowledge base[D]. Changchun: Jilin University, 2021.] DOI: 10.27162/d.cnki.gjlin.2021.001257.

9 王桂彬，庞博 . 名老中医隐性知识发现与医案解构模式研究 [J]. 中华中医药杂志，2023, 38(5): 2230–2234. [Wang GB, Pang B. Research on the model of tacit knowledge discovery and medical case deconstruction of famous old Chinese medicine experts[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2023, 38(5): 2230–2234.] https://www.zhangqiaokeyan.com/academic–journal–cn_detail_thesis/02012107144540.html.

10 张婷婷，王亚强，蒋艺宁，等 . 构建中医辨证解释体系的挑战与思路 [J]. 中医杂志，2024, 65(5): 445–448, 454. [Zhang TT, Wang YQ, Jiang YN, et al. Challenges and ideas for constructing a TCM syndrome differentiation and interpretation system[J]. Journal of Traditional Chinese Medicine, 2024, 65(5): 445–448, 454.] DOI: 10.13288/j.11–2166/r.2024.05.001.

11 王一名，代欣玥，郭曼萍，等 . 中医药真实世界研究数据转化方法 [J]. 中国循证医学杂志，2023, 23(9): 1081–1088. [Wang YM, Dai XY, Guo MP, et al. Data transformation method of real world study on traditional Chinese medicine[J]. Chinese Journal of Evidence–Based Medicine, 2023, 23(9): 1081–1088.] DOI: 10.7507/1672–2531.202302094.

12 王志国，李思婷 . 关于中医病名、证候、症状、体征、

病状、临床表现等术语规范化 [J]. 中医学 , 2021, 10(6): 5. [Wang ZG, Li ST. Standardization of TCM terms such as disease names, syndromes, symptoms, signs, symptoms and clinical manifestations[J]. Traditional Chinese Medicine, 2021, 10(6): 5.] DOI: 10.12677/TCM.2021.106105.

13 仝小林 , 房敏 , 高慧 , 等 . 2021 年度中医药重大科学问题和工程技术难题 [J]. 中医杂志 , 2021, 62(11): 921–929. [Tong XL, Fang M, Gao H, et al. Major scientific issues and engineering and technical problems of traditional Chinese medicine in 2021[J]. Journal of Traditional Chinese Medicine, 2021, 62(11): 921–929.] DOI: 10.13288/j.11–2166/r.2021.11.001.

14 潘越 , 侯胜田 , 赵曙光 , 等 . 北京中医药文化传播发展报告 (2015)[M]. 北京 : 社会科学文献出版社 , 2015. [Pan Y, Hou ST, Zhao SG, et al. Report on TCM culture communication development of Beijing (2015)[M]. Beijing: Social Sciences Academic Press. 2015.]

15 杨佳澄 . 基于群体智能的中医辨证诊断研究 [D]. 兰州：兰州交通大学 , 2019. [Yang JC. Research on TCM diagnosis based on swarm intelligence[D]. Lanzhou: Lanzhou Jiaotong University, 2019.] DOI: 10.27205/d.cnki.gltec.2019.000011.

16 梁洁 , 和思敏 , 陈淑婷 , 等 . 纵向研究中控制时依混杂的 G 方法 [J]. 中华流行病学杂志 , 2021, 42(10): 5. [Liang J, He SM, Chen ST, et al. In longitudinal studies, the control method was conflated according to the G method[J]. Chinese Journal of Epidemiology, 2021, 42(10): 5.] DOI: 10.3760/cma.j.cn112338–20200731–01001.

17 张佳康 . 基于数据挖掘自拟消瘿汤治疗气郁痰阻型甲状腺结节的临床观察 [D]. 哈尔滨：黑龙江中医药大学, 2023. [Zhang JK. Clinical observation of self–made Xiaoying decoction based on data mining in the treatment of thyroid nodules with qi stagnation and phlegm obstruction[D]. Harbin: Heilongjiang University of Chinese Medicine, 2023.] DOI: 10.27127/d.cnki.ghlzu.2023.000374.

18 庄逸洋 , 郑升鹏 , 陈文嘉 , 等 . 国医大师邓铁涛治疗冠心病用药规律的数据挖掘研究 [J]. 时珍国医国药, 2016, 27(12): 3025–3027. [Zhuang YY, Zheng SP, Chen WJ, et al. Data mining research on the medication rules of Deng Tietao, a master of traditional Chinese medicine, in the treatment of coronary heart disease[J]. Lishizhen Medicine and Materia Medica Research, 2016, 27(12): 3025–3027.] DOI: 10.3969/j.issn.1008–0805.2016.12.071.

19 刘静 , 杨建新 , 王春晓 , 等 . 基于《伤寒论》六经辨证体系的腰椎间盘突出症中医证型规律研究 [J]. 中国中医骨伤科杂志 , 2023, 31(7): 12–16. [Liu J, Yang JX, Wang CX, et al. Study on the law of TCM syndromes of lumbar disc herniation based on the syndrome differentiation system of six meridians in treatise on febrile diseases[J].Chinese Journal of Traditional Medical Traumatology & Orthopedics, 2023, 31(7): 12–16.] DOI: 10.20085/j.cnki.issn1005–0205.230703.

20 李毅 , 刘艳 , 刘力 , 等 . 溃疡性结肠炎中医症状学主成分分析 [J]. 中医药导报 , 2016, 22(7): 32–35. [Li Y, Liu Y, Liu L, et al. TCM Symptoms of ulcerative colitis of principal component analysis[J]. Guiding Journal of Traditional Chinese Medicine and Pharmacy, 2016, 22(7): 32–35.] DOI: 10.13862/j.cnki.cn43–1446/r.2016.07.010.

21 谭楠楠 , 章轶立 , 杜康佳 , 等 . 基于主成分分析的慢性心力衰竭中医症状与证候研究 [J]. 中华中医药杂志 , 2021, 36(7): 4265–4267. [Tan NN, Zhang YL, Du KJ, et al. Exploration of TCM symptoms and syndromes of chronic heart failure based on principal component analysis[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2021, 36(7): 4265–4267.] https://d.wanfangdata.com.cn/periodical/ChlQZXJpb2RpY2FsQ0hJTmV3UzIwMjMxMjI2Eg96Z3l5eGIyMDIxMDcxMTcaCGg3bW8xdGQ2.

22 徐小港 , 王钰 , 徐义峰 , 等 . 基于数据挖掘的痛经用药规律研究 [J]. 西部中医药 , 2024, 37(2): 99–103. [Xu XG, Wang Y, Xu YF, et al. Study on medication rule of dysmenorrhea based on data mining[J]. Western Journal of Traditional Chinese Medicine, 2024, 37(2): 99–103.] DOI: 10.12174/j.issn.2096–9600.2024.02.19.

23 史媛媛 . 基于明清时期及近现代医案分析的不寐证治规律研究 [D]. 沈阳：辽宁中医药大学 , 2022. [Shi YY. Based on the analysis of medical cases in the Ming and Qing dynasties and modern times, the law of insomnia syndrome treatment is studied[D]. Shengyan: Liaoning University of Traditional Chinese Medicine, 2022.] DOI: 10.27213/d.cnki.glnzc.2022.000095.

24 段一凡 , 唐明坤 , 孙海霞 , 等 . 面向疾病风险智能预测研究过程的电子病历数据质量需求模型构建 [J]. 中国循证医学杂志 , 2023, 23(9): 1072–1080. [Duan YF,

Tang MK, Sun HX, et al.Exploring data quality for machine learning-based disease risk predictions with electronic medical records[J]. Chinese Journal of Evidence-Based Medicine, 2023, 23(9): 1072-1080.] DOI: 10.7507/1672-2531.202301076.

25 余学杰，李书珍，李晓燕，等．基于决策树提取中医专家辨证规律初探[J]. 辽宁中医杂志，2015, 42(1): 19-24. [Yu XJ, Li SZ, Li XY, et al. Study on extracting Chinese medicine experts' diagnostic rules by decision tree[J].Liaoning Journal of Traditional Chinese Medicine,2015, 42(1): 19-24.] DOI: 10.13192/j.issn.1000-1719.2015.01.006.

26 刘广，孙艳秋，裴媛．基于 C4.5 决策树算法的中医胃炎实验数据分类挖掘研究[J]. 中华中医药学刊，2016, 34(12): 2958-2961. [Liu G, Sun YQ, Pei Y, et al. Classified mining of TCM gastritis based on C4.5 decision tree algorithm[J]. Study Journal of Traditional Chinese Medicine, 2016, 34(12): 2958-2961.] DOI: 10.13193/j.issn.1673-7717.2016.12.039.

27 黄嘉韵，郭宏，邝艳萍．基于决策树算法的鼻鼽辨证规律初步研究[J]. 中华中医药杂志，2016, 31(11): 4770-4773. [Huang JY, Guo H, Kuang YP. Preliminary research on regularity of syndrome differentiation of allergic rhinitis based on decision tree algorithm[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2016, 31(11): 4770-4773.] DOI: CNKI:SUN:BXYY.0.2016-11-114.

28 苏翀，任曈，王国品，等．利用决策树建立慢性阻塞性肺病中医诊断模型[J]. 计算机工程与应用，2019, 55(3): 225-230. [Su C, Ren T, Wang GP, et al. Using K-L divergence based decision tree to build traditional Chinese medicine diagnosis model on COPD[J]. Computer Engineering and Applications, 2019, 55(3): 225-230.] DOI: 10.3778/j.issn.1002-8331.1710-0089.

29 宫文浩，兰天莹，杨燕，等．基于随机森林和偏相关分析的小儿肺炎痰热闭肺证中医证候诊断模型研究[J]. 中华中医药杂志，2023, 38(9): 4497-4501. [Gong WH, Lan TY, Yang Y, et al. Research on traditional Chinese medicine syndrome diagnosis model of pediatric pneumonia with accumulation of phlegm-heat syndrome based on random forest and partial correlation analysis[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2023, 38(9): 4497-4501.] https://www.cnki.com.cn/Article/CJFDTOTAL-BXYY202309096.htm.

30 姜超，冯哲，王均琴，等．三种机器学习方法在脑出血中医辨证分类中应用的比较研究[J]. 中国卫生统计，2023, 40(6): 921-924, 928. [Jiang C, Feng Z, Wang JQ, et al. A comparative study on the application of three machine learning methods in TCM syndrome differentiation and classification of intracerebral hemorrhage[J]. Chinese Journal of Health Statistics, 2023, 40(6): 921-924, 928.] DOI: 10.11783/j.issn. 102-3674.2023.06.028.

31 朱文锋，朱咏华，黄碧群．采用贝叶斯网络运算进行中医辨证的探讨[J]. 广州中医药大学学报，2006, 23(6): 4. [Zhu WF, Zhu YH, Huang BQ. Bayesian network computing was used to discuss syndrome differentiation in TCM[J]. Journal of Guangzhou University of Traditional Chinese Medicine, 2006, 23(6):4.] DOI: 10.3969/j.issn.1007-3213.2006.06.001.

32 刘鑫子，李自艳，郑思思，等．基于聚类分析及贝叶斯网络的双相情感障碍抑郁发作中医证候的横断面研究[J]. 中医杂志，2024, 65(1): 79-85. [Liu XZ, Li ZY, Zheng SS, et al. A cross-sectional study of TCM patterns of depressive episodes in bipolar disorder based on cluster analysis and Bayesian network[J]. Journal of Traditional Chinese Medicine, 2024, 65(1): 79-85.] DOI: 10.13288/j.11-2166/r.2024.01.015.

33 甘小金，陈艳，马秀丽．基于贝叶斯网络的王子瑜教授治疗子宫内膜异位症的辨证规律研究[J]. 世界中西医结合杂志，2019, 14(10): 3. [Gan XJ, Chen Y, Ma XL. Study on professor Wang Ziyu's TCM syndromes in the treatment of endometriosis based on Bayesian networks[J]. World Journal of Integrated Traditional and Western Medicine, 2019, 14(10): 3.] DOI: CNKI:SUN:SJZX.0.2019-10-006.

34 任冷，周维民．针对非平衡多分类问题 SVM 算法的优化研究与应用[J]. 电脑知识与技术，2016, 12(5): 218-220. [Ren L, Zhou WM. Optimization research and application of SVM algorithm for unbalanced multi-classification problem[J]. Computer Knowledge and Technology, 2016, 12(5): 218-220.] DOI: 10.14004/j.cnki.ckt.2016.0618.

35 许明东，马晓聪，温宗良，等．支持向量机在高血压病中医证候诊断中的应用[J]. 中华中医药杂志，2017, 32(6): 2497-2500. [Xu MD, Ma XQ, Wen ZL, et al. Application of support vector machine in the diagnosis of hypertension in TCM syndrome[J].China Journal of

Traditional Chinese Medicine and Pharmacy, 2017, 32(6): 2497–2500.] DOI: CNKI:SUN:BXYY.0.2017–06–044.

36 顾天宇, 严壮志, 蒋皆恢. 基于支持向量机的中风病中医证候分类[J]. 中医药信息, 2021, 38(9): 1–3. [Gu TY, Yan ZZ, Jiang JH. Classification of TCM syndrome patterns of stroke based on SVM[J]. Information on Traditional Chinese Medicine, 2021,38(9): 1–3.] DOI: 10.19656/j.cnki.1002–2406.20210901.

37 王伟杰, 唐晓颇, 王新昌, 等. 基于临床辨证的类风湿关节炎常见中医证候 Logistic 回归分析[J]. 中华中医药杂志, 2019, 34(2): 807–810. [Wang WJ, Tang XP, Wang XC, et al. Logistic regression analysis on TCM syndromes of rheumatoid arthritis based on syndrome differentiation[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2019, 34(2): 807–810.] DOI: CNKI:SUN:BXYY.0.2019–02–100.

38 仲芳. 基于量效分析的《吴鞠通医案》数据挖掘研究[D]. 上海: 上海中医药大学, 2020. [Zhong F, Data mining research on the case of Wu Jutong based on quantum–effect analysis[D]. Shanghai: Shanghai University of Traditional Chinese Medicine, 2020.] DOI: 10.27320/d.cnki.gszyu.2020.000551.

39 孟光磊, 丛泽林, 宋彬, 等. 贝叶斯网络结构学习综述[J/OL]. 北京航空航天大学学报, 1–24 [2024–03–05]. [Meng GL, Cong ZL, Song B, et al. Review of Bayesian network structure learning[J/OL]. Journal of Beijing University of Aeronautics and Astronautics, 1–24 [2024–03–05].] https://doi.org/10.13700/j.bh.1001–5965.2023.0445.

40 Brown TB, Mann B, Ryder N, et al. Language models are few–shot learners[J]. 2020. DOI: 10.48550/arXiv.2005.14165.

41 Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. 2022. DOI: 10.48550/arXiv.2202.03629.

42 尹玉洁, 常丽萍, 朱垚, 等. 基于脉络学说指导和医案数据挖掘的心血管事件链证治规律分析[J]. 中国实验方剂学杂志, 2022, 28(18): 144–151. [Yin YJ, Chang LP, Zhu Y, et al. Vessel–collateral theory guided rules of syndrome and treatment of cardiovascular event chain based on medical record data mining and analysis[J]. Chinese Journal of Experimental Traditional Medical Formulae, 2022, 28(18): 144–151.] DOI: 10.13422/j.cnki.syfjx.20220651.

收稿日期：2023 年 12 月 26 日 修回日期：2024 年 03 月 12 日
本文编辑：桂裕亮 曹 越

引用本文：夏鑫, 牟玮, 李艳芬, 等. 基于机器学习技术挖掘中医名家医案数据的方法探讨[J]. 医学新知, 2024, 34(4): 448–457. DOI: 10.12173/j.issn.1004–5511.202312129
Xia X, Mu W, Li YF, et al. Approaches to the mining of traditional Chinese medical experts' case histories using machine learning techniques[J]. Yixue Xinzhi Zazhi, 2024, 34(4): 448–457. DOI: 10.12173/j.issn.1004–5511.202312129